GUÍA DE EJERCITACIÓN N° 1

Regresión Lineal

1) Trabajamos con los datos "phones" del paquete MASS en RStudio, que representan las llamadas telefónicas de Bélgica 1950-1973, donde la variable independiente (x) representan los dos últimos dígitos de los años comprendidos entre 1950-1973 y la variable respuesta (y) es el número de llamadas telefónicas (millones) realizadas en Bélgica.

a) Identifique los datos y plantee un modelo que permita predecir las llamadas telefónicas en función de los años.
b) Grafique los datos para el modelo planteado.
c) Plantee el modelo de regresión y escriba su ecuación.
d) Realice el diagnóstico del modelo.
e) Calcule y grafique los intervalos de confianza para la respuesta media y para las predicciones.
f) Determine cuál fue el número de llamadas telefónicas a mitad del año 1965.

2) Utilizamos los datos "Boston" del paquete MASS, que recolecta la información de la mediana del valor de la vivienda en 506 áreas residenciales de Boston. Junto con el precio, se han registrado 13 variables adicionales:

crim: ratio de criminalidad per capita de cada ciudad.

zn: proporción de zonas residenciales con edificaciones de más de 25.000 pies cuadrados.

indus: proporción de zona industrializada.

chas: si hay río en la ciudad (= 1 si hay río; 0 no hay).

nox: concentración de óxidos de nitrógeno (partes per 10 millón).

rm: promedio de habitaciones por vivienda.

age: proporción de viviendas ocupadas por el propietario construidas antes de 1940.

dis: media ponderada de la distancias a cinco centros de empleo de Boston.

rad: índice de accesibilidad a las autopistas radiales.

tax: tasa de impuesto a la propiedad en unidades de $10,000.
ptratio: ratio de alumnos/profesor por ciudad.
black: 1000(Bk - 0.63)^2 donde Bk es la proporción de gente de color por ciudad.
lstat: porcentaje de población en condición de pobreza.
medv: valor mediano de las casas ocupadas por el dueño en unidades de $1000s.
Analice la relación entre el valor mediano de las casas y el porcentaje de población en condición de pobreza.

a) Identifique y analice el conjunto de datos.
b) Grafique los datos para las variables en estudio.
c) Plantee el modelo de regresión y escriba su ecuación.
d) Grafique y extraiga conclusiones de la tabla de ANOVA.
e) Realice el diagnóstico del modelo utilizando los residuos, extraiga conclusiones.
f) Calcule los intervalos de confianza para la respuesta media del valor de las casas y para las predicciones.
g) Determine el valor promedio de todas las viviendas que se encuentran en una población para un valor de lstat = 10

3) Utilizaremos los datos marketing del paquete datarium, que resume el número de ventas de un producto determinado en relación con el presupuesto invertido en publicidad en Youtube, Facebook y en el periódico. Se analizará la relación entre las ventas (sales) y el presupuesto en euros en publicidad en Youtube (youtube).

a) Identifique y analice el conjunto de datos.
b) Grafique los datos para las variables en estudio.
c) Plantee el modelo de regresión y escriba su ecuación.
d) Grafique y extraiga conclusiones de la tabla de ANOVA.
e) Realice el diagnóstico del modelo utilizando los residuos, extraiga conclusiones.
f) Determine el número de ventas que tendremos si invertimos 10, 100 o 300 mil dólares en publicidad en youtube.
g) Calcule y grafique los intervalos de confianza para la respuesta media y para las predicciones.

4) Retomamos los datos del ejercicio anterior y analizamos la relación de todas sus variables.

a) Grafique los datos de las tres variables para analizar la relación entre ellas.
b) Plantee el modelo de regresión y escriba su ecuación.
c) Determine el error estándar residual (RSE) y la tasa de error.
d) Calcule los intervalos de confianza.
e) Utilice los métodos basados en criterios de información.
f) Plantee y extraiga conclusiones de la tabla de ANOVA.
g) Aplique los métodos de selección por pasos.
h) Evalúe la multicolinealidad del modelo.
i) Realice el diagnóstico del modelo utilizando los residuos, extraiga conclusiones.

5) Del ejercicio 2 se desea generar un modelo que permita explicar el precio de la vivienda de una población empleando para ello cualquiera de las variables disponibles en el dataset Boston y que resulten útiles en el modelo.

a) Plantee el modelo de regresión y escriba su ecuación.
b) Determine el error estandar residual (RSE) y la tasa de error.
c) Calcule los intervalos de confianza.
d) Aplique los métodos de selección de predictores por pasos.
e) Utilice los métodos basados en criterios de información.
f) Plantee y extraiga conclusiones de la tabla de ANOVA.
g) Evalúe la multicolinealidad del modelo.
h) Realice el diagnóstico del modelo utilizando los residuos, extraiga conclusiones.

6) Se quiere predecir el número de bicicletas que se van a alquilar (Variable *cnt*), para lo cual se debe preparar el conjunto de los datos para crear un modelo de regresión. Se proporciona un conjunto de datos con 731 observaciones y 10 variables, donde la variable *cnt* es la variable a predecir. Realice los siguientes pasos de las fases de visualización y pre-procesado de datos:

a) Cargar los datos y transformar aquellas variables que se consideren factor.

b) Dividir el conjunto de datos en un conjunto de entrenamiento y otro de validación (80%-20% de las observaciones).
c) Centrar y escalar las variables numéricas.
d) Crear variables binarias (dummys).
e) Plantear el modelo de regresión y escriba su ecuación.
f) Aplique los métodos de selección de predictores por pasos.
g) Plantee y extraiga conclusiones de la tabla de ANOVA.
h) Evalúe la multicolinealidad del modelo.
i) Realice el diagnóstico del modelo utilizando los residuos, extraiga conclusiones.

7) Del conjunto de datos Salaries del paquete car, que contiene los datos del salario académico durante nueve meses del 2008-2009 para Profesores adjuntos, Profesores asociados y Profesores titulares, en una universidad de los EE.UU.. Los datos se recopilaron como parte del esfuerzo continuo de la administración de la universidad para monitorear las diferencias salariales entre hombres y mujeres que trabajan en la facultad.

a) Identifique y analice el conjunto de datos.
b) Transforme la variable categórica sexo en una variable indicadora (o dummy).
c) Plantee el modelo de regresión y escriba su ecuación.
d) Evalúe la contribución global del modelo y extraiga conclusiones. (Prueba F y t)
e) Realice el mismo análisis anterior al resto de las variables categóricas.

8) Utilice el set de datos de "Prestige", en el cual se registra los resultados de un estudio realizado en Canadá sobre el prestigio de las profesiones.

education: Educación media de los titulares ocupacionales.

income: Ingreso promedio en dólares.

women: Porcentaje de mujeres por ocupación.

Prestige: Prestigio de la ocupación, resultado de una encuesta social realizada a mediados de la década de 1960.

census: Código ocupacional del censo canadiense.

type: Tipo de ocupación. Un factor con niveles: bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

a) Analice e interprete el conjunto de datos.
b) Transforme las variable dummy
c) Implemente el modelo de regresión lineal múltiple y escriba su ecuación.
d) En base al modelo obtenido realice la predicción con los siguientes datos:

educación = 11.5; income = 6500; women = 14.5; census = 5135; type = prof

e) Aplique los metodos de selección de predictores por pasos.